A Beginners Guide to Phred, Phrap and Consed

Training offered by the Advanced Biomedical Computing Center, NCI/Frederick

Beena Neelam
email:bneelam@ncifcrf.gov
call: 301 846 5779

Objectives

- **Objectives of the course**
  - What are phred, phrap and consed?
  - prepare your chromat files for running phred and phrap on the ABCC UNIX machines
  - run phred- base calling program
  - run phrap- assembly program
  - view contigs using consed- a visualization tool

What are Phred/Phrap/Consed?

- **Phred/Phrap/Consed is a distributed package, free for academic use, for:**
  - Reading trace files (chromatograms)
  - Quality value assignment to each base
  - Vector and repeat sequences masking
  - Sequence assembly
  - Assembly visualization and editing
  - Automatic finishing

Phred

- **Developed at the University of Washington**
  - Brent Ewing, et. al., Base-calling of automated sequencer traces using phred. I. Accuracy assessment. 1998. Genome Research 8:175-185.
  - Brent Ewing and Phil Green, Base-calling of automated sequencer traces using phred. II. Error probabilities. 1998. Genome Research 8:186-194.

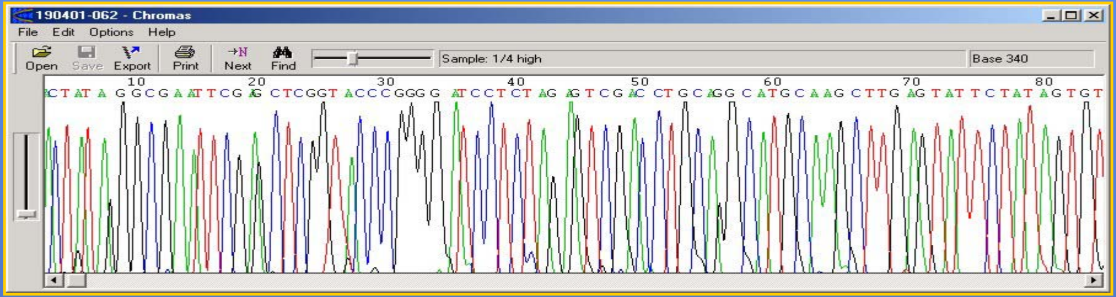
  – Reads DNA sequencer trace data from chromatogram in SCF (standard chromatogram format), ABI (373/377/3700), ESD (MegaBACE) and LI-COR

  – calls bases (sequencing machine and chemistry specific)

contd.

Trace File Quality

High

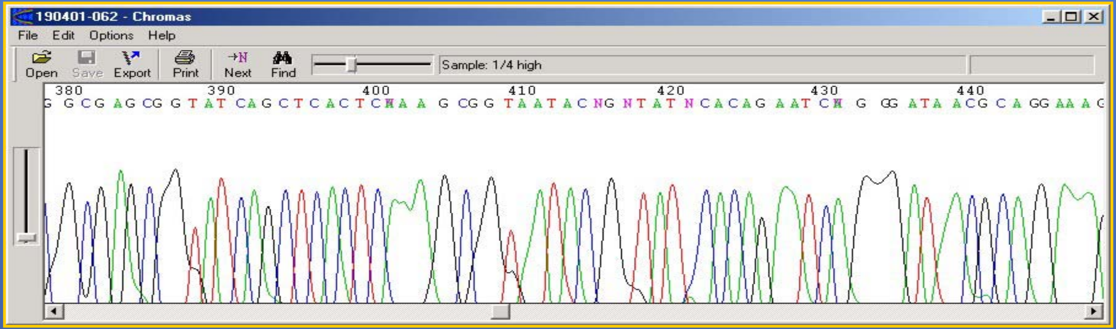


Sample: 1/4 high

Base 340

```
       10          20             30              40            50              60              70             80
CTAT A GGCGAATTCGAG CTCGGT ACCCGGG G ATCCTCT AG AGTCGAC CTGCAGG CATGCAAG CTTG AG TAT TCTAT AGTGT
```

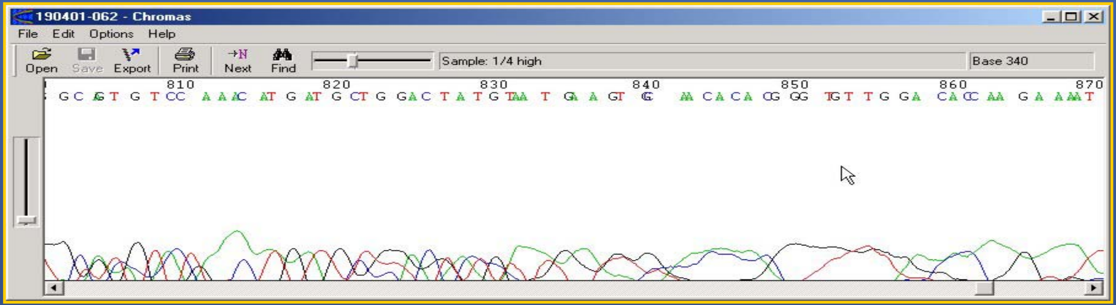
Medium



Sample: 1/4 high

```
380          390            400             410            420              430             440
G GCG AG CG G TAT CAG CTCACTCNAA G CGG TAATACNG NTAT NCACAG AAT CN G GG ATA ACG CA GG AA A G
```

Low



Sample: 1/4 high

Base 340

```
          810            820              830              840             850              860            870
GC AGT G TCC AAAC AT G AT G CT G GAC TATG TAA T GA A GT G AA CACA CG GG TGT TG GA CA CC AA G A AAT
```

Phred

- assigns quality score to bases (related to base call error probability)

- Determines file format automatically

- Writes sequences to output in FASTA format or PHD format or SCF (Sequence Comparison Interchange Format) format

- Outputs quality scores for each trace file

- Optionally outputs Polyphred values file

phd2fasta

- **phd2fasta program**
  - – converts .phd files to sequence in multifasta format
  - – writes .qual file (quality scores) for each trace file
- **Output:**
  - – fasta.seq contains fasta sequences
  - – fasta.seq.qual contains quality scores

Crossmatch

- A general purpose utility for comparing two DNA sequences-efficient Smith Waterman algorithm

- More sensitive than BLASTN but slower, allows gaps.

- Can be used to
  - compare a set of DNA sequences to a set of vector sequences
  - compare contigs derived by two methods of assembly
  - align contigs to reference sequence

Phrap

- **F(ph)ragment Assembly Program- Phil Green**
  - A program for assembling DNA sequence data-reads of 500-700 bp,
  - Uses entire read, uses user supplied and internally computed quality values
  - Constructs contigs as a mosaic of highest quality reads not as consensus

Phrap

- –gives extensive information about the assembly -easier to troubleshoot
  - –in <filename>.phrap.out file
    - – Residue counts, Read name analysis, Input quality, Regions converted to N's
- –can handle large datasets

Consed

- **A program for viewing and editing sequence assemblies** -David Gordon
  - Gordon, D., C. Desmarais, and P. Green. 2001. Automated Finishing with Autofinish. Genome Research. 11(4):614-625.

  - The version we are using is consed13.0
  - View contigs assembled with Phrap, view single or multiple trace
  - Add reads and reassemble
  - Several useful and timesaving navigation features to view aligned reads, low/high quality regions
  - Autofinish: Close gaps, improve sequence quality, determine the relative orientation of contigs
  - several other features...

Running the programs

- **Two ways to run the programs**
  - Run as a single program called PhredPhrap
    - PhredPhrap program runs Phred, runs phd2fasta to make the fasta format sequence file from the .phd files, runs cross match to mask vector and then runs phrap to assemble the reads
  - Run sequentially
    - Run Phred, create .phd files, then run phd2fasta (providing quality scores if necessary), then run crossmatch to mask vector and then run phrap

Running Phred (base calling)

- **Make directories**
  - chromat_dir, phd_dir and edit_dir and place chromatogram files in chromat_dir
- **Run phred**
  - either compressed file or all the individual trace files
  - phd_dir will have .phd files
- **Run phd2fasta**
  - with .phd files as input, and edit_dir as output dir. Result is a set of fasta files in edit_dir

phredpar.dat file

- **phredpar.dat file**
  - is a chemistry parameter file, contains dye primer data

- **Location of phredpar.dat file:**
  - ncisgi: /abcc/apps/IRIX64/phred02/

Example of a phredpar.dat file:

```
###########################################################################
# phredpar.dat - phred parameter file: 020425
#  known chemistries: primer, terminator, unknown
#  known dyes       : rhodamine, d-rhodamine, big-dye  energy-transfer, bodipy, unknown
#  known machines   : ABI_373_377, MolDyn_MegaBACE, ABI_3700, LI-COR_4000,
#  Beckman_CEQ_2000, ABI_3100
# Notes:
#       (1) enclose the `dye primer' name in double quotes and include spaces in the names.
#       (2) leave one or more spaces between the `dye primer' and chemistry names, between
#       the chemistry and dye names, and between the dye and machine names.
#       (3) add entries between the `begin chem_list' and  `end chem_list' lines.
#       (4) `dye primer' string `__phred_default__'  identifies the chemistry, dye, and machine
#       used  when the
#       chromatogram dye primer string is  missing from either the chromatogram or this file and
#       phred is run with the -process_nomatch  option.
##########################################################################
begin chem_list
""                          primer        big-dye              ABI_3700
"DP6%25Ac{-21M13}"             primer        rhodamine             ABI_373_377
"DP6%Ac{-21M13}"               primer        rhodamine             ABI_373_377
"DP6%25Ac{M13Rev}"             primer        rhodamine             ABI_373_377
```

Setting up to run the programs

- **ABCC UNIX account**
  - – http://www2.ncifcrf.gov/
- **Transferring trace files**
  - – Finch
  - – ftp from desktop
  - – as compressed file on CD
  - – already on UNIX machine
- **Edit your .cshrc file:**
  - – add line source /abcc/apps/perl/setup.csh

phred, phrap and crossmatch location

Phred on ncisgi:

/abcc/apps/IRIX64/phred02/phred

Phrap on ncisgi:

/abcc/apps/IRIX64/phrap02/phrap

cross_match on ncisgi:

/abcc/apps/IRIX64/phrap02/cross_match

Preparing trace files for analysis

- %cd tar_folder
- %gunzip ab1_F.tar.gz
- %tar –xf ab1_F.tar
- -xf, extract (x) files (f) from an archive
- Result: list of .ab1 files

Running phred

- **Make subdirectories**
  - in the directory containing the compressed file:
  - make chromat_dir phd_dir edit_dir
- **Run phred**
  - %phred –id chromat_dir –pd phd_dir
- **Output**
  - chromat_dir has the chromat files eg *.ab1 format
  - phd_dir will have corresponding *.phd files created by phred

Phred options

- **Command line options**
  - options are preceded by a '-'
  - input options, processing options and output options
  - input options: -id, -if
  - processing options: -exit_nomatch
  - output options: -sa, -qa, -sd, -qd, -pd

Running phd2fasta

- **phd2fasta program**
  - converts .phd files to sequence in multifasta format
  - writes .qual file (quality scores) for each trace file
  - %phd2fasta –id phd_dir –os edit_dir/fasta.seq –oq edit_dir/fasta.seq.qual
- **Output:**
  - fasta.seq contains fasta sequences
  - fasta.seq.qual contains quality scores

An example of a .phd file

- BEGIN_SEQUENCE 522e10_R_N1-9H12.g
- BEGIN_COMMENT
- CHROMAT_FILE: 522e10_R_N1-9H12.g
- ABI_THUMBPRINT: 0
- PHRED_VERSION: 0.020425.c
- CALL_METHOD: phred
- QUALITY_LEVELS: 99
- TIME: Fri Jan 23 11:41:30 2004
- TRACE_ARRAY_MIN_INDEX: 0
- TRACE_ARRAY_MAX_INDEX: 14306
- TRIM: 23 878 0.0500
- TRACE_PEAK_AREA_RATIO: 0.0091
- CHEM: term
- DYE: big
- END_COMMENT
- BEGIN_DNA
- g 6 7
- g 6 16
- a 6 37
- t 6 40
- t 8 60
- g 3 71
- c 3 80

Running crossmatch to screen vector

- **Run crossmatch**

  -
    ```
    %cross  fasta.seq
    /abcc/apps/IRIX64/phrap02/allvector.seq
    –minmatch 10 –minscore 20 –screen > screen.out
    ```
- **The –screen option**
  - causes files called fasta.seq.screen to be created, contains vector masked sequences.
- **screen.out**
  - information about screen parameters and vector matches found

Running phrap

- phrap = F(ph)ragment assembly program or Phil’s revised assemble program

-
  ```
  %phrap –new_ace edit_dir/fasta.seq > phrap.out
  ```

- All the seqs. to be assembled are typically given in a single input file.

- Input files for Phrap: Sequence file (typically single multifasta file) and Quality file, both fasta format

- for each input fasta.seq file if a corresponding fasta file containing data quality information is provided, it greatly improves the accuracy of assembly.

- if C05D11.reads.screen is your input file corresponding quality file should be C05D11.reads.screen.qual

Phrap options

- Scoring of pair wise alignments
- Banded search
- Filtering of matches
- Input data interpretation
- Assembly
- Consensus sequence construction
- Output

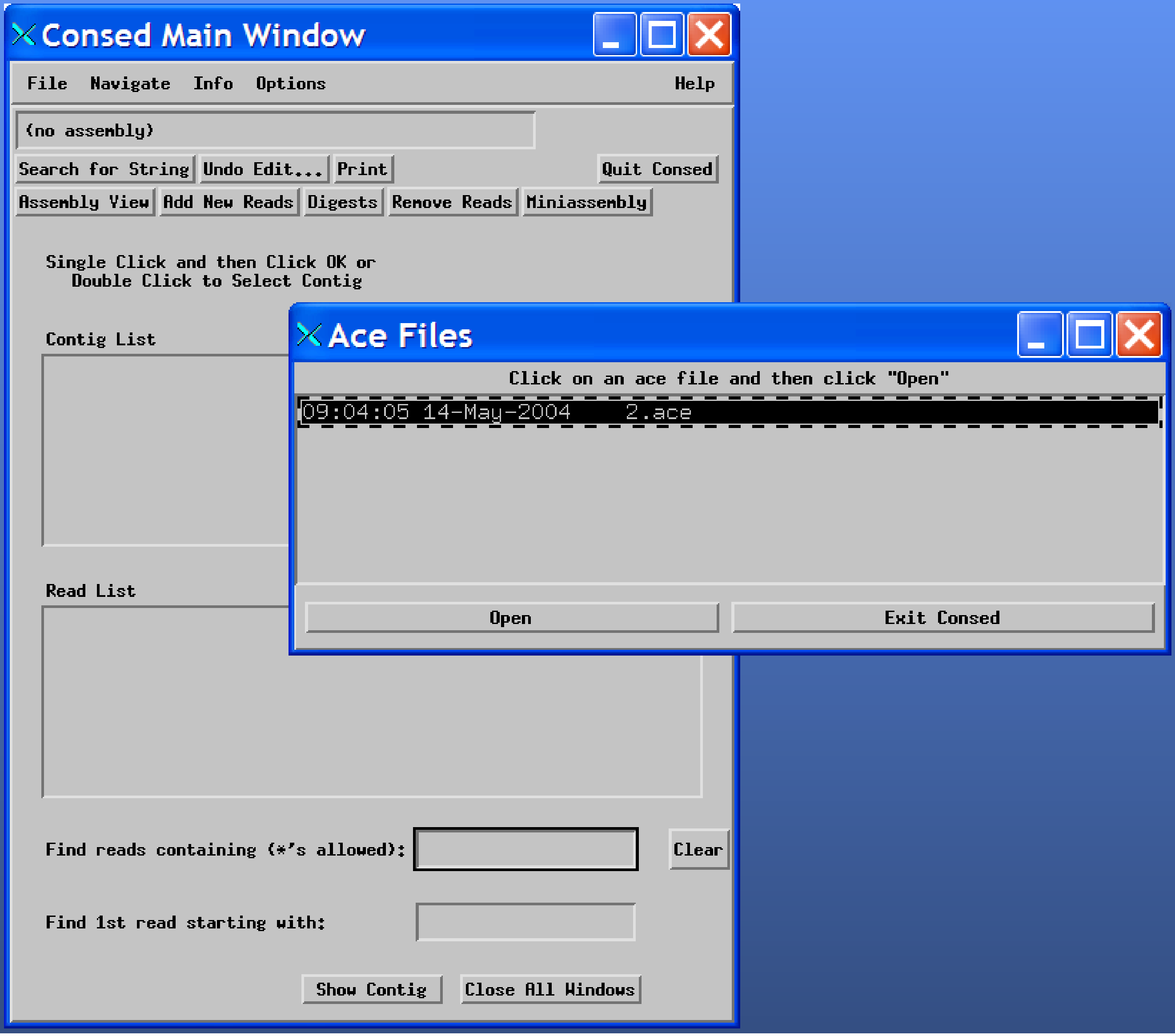
Phrap output files

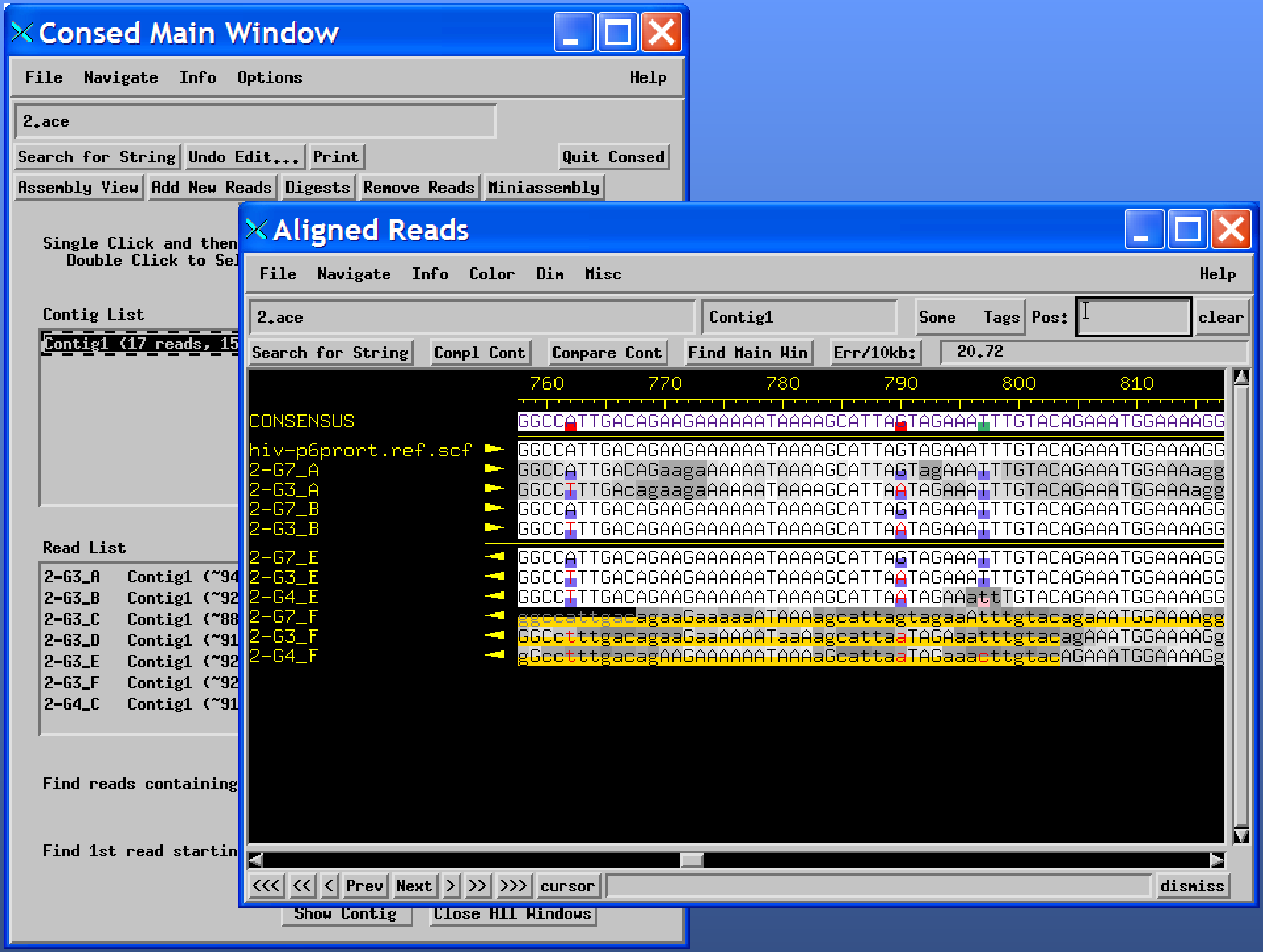
- *.contigs – fasta file containing the contigs
  - Contigs with more than one read
  - Singletons (single reads with a match to some other contig but that couldn’t be merged consistently to it)
- *.singlets – fasta file of the singlet reads
  - Reads with no match to other read
- *.ace
  - allows for viewing the assembly using Consed
- *.view
  - required for viewing the assembly using Phrapview

Running Consed

- Viewing and editing sequence assemblies with Consed
  - %cd edit_dir
  - Type consed13

Consed windows





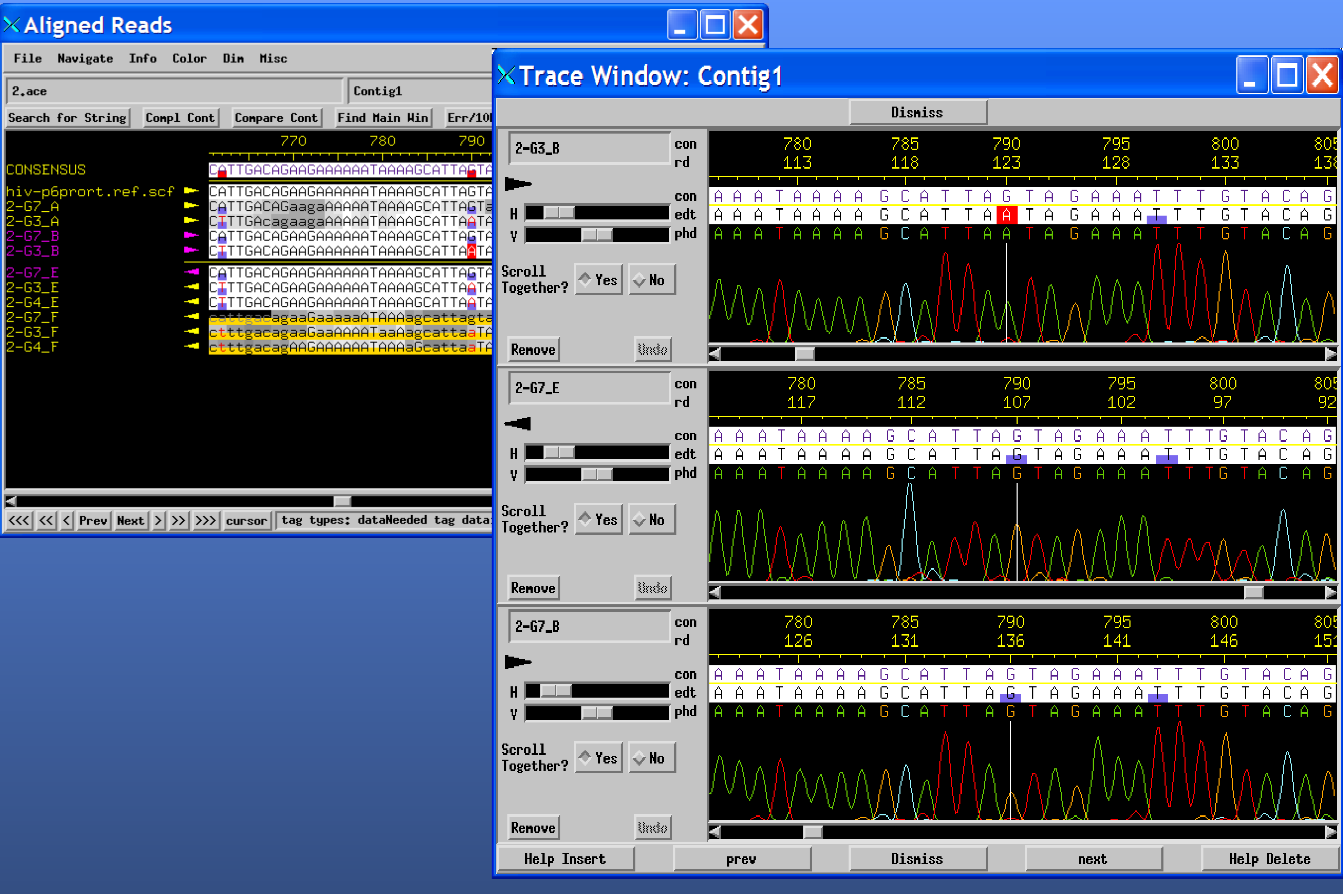
```
2.ace
Contig1 (17 reads, 15

2-G3_A    Contig1 (~94
2-G3_B    Contig1 (~92
2-G3_C    Contig1 (~88
2-G3_D    Contig1 (~91
2-G3_E    Contig1 (~92
2-G3_F    Contig1 (~92
2-G4_C    Contig1 (~91
```

```
2.ace                                   Contig1
Err/10kb: 20.72

                       760       770       780       790       800       810
CONSENSUS              GGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAAATGGAAAAGG
hiv-p6prort.ref.scf  ► GGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAAATGGAAAAGG
2-G7_A               ► GGCCATTGACAGaagaAAAAATAAAAGCATTAGTagAAATTTGTACAGAAATGGAAAagg
2-G3_A               ► GGCCTTTGAcagaagaAAAAATAAAAGCATTAATAGAAATTTGTACAGAAATGGAAAagg
2-G7_B               ► GGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAAATGGAAAAGG
2-G3_B               ► GGCCTTTGACAGAAGAAAAAATAAAAGCATTAATAGAAATTTGTACAGAAATGGAAAAGG

2-G7_E               ◄ GGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAAATGGAAAAGG
2-G3_E               ◄ GGCCTTTGACAGAAGAAAAAATAAAAGCATTAATAGAAATTTGTACAGAAATGGAAAAGG
2-G4_E               ◄ GGCCTTTGACAGAAGAAAAAATAAAAGCATTAATAGAAattTGTACAGAAATGGAAAAGG
2-G7_F               ◄ ggccattgacagaaGaaaaaATAAAagcattagtagaaAtttgtacagaAATGGAAAAgg
2-G3_F               ◄ GGCctttgacagaaGaaAAAATaaAagcattaaTAGAaatttgtacagAAATGGAAAAGg
2-G4_F               ◄ gGccttttgacagAAGAAAAAATAAAaGcattaaTAGaaacttgtacAGAAATGGAAAAGg
```



```
                                770       780       790
CONSENSUS                       CATTGACAGAAGAAAAAATAAAAGCATTAGTA
hiv-p6prort.ref.scf             CATTGACAGAAGAAAAAATAAAAGCATTAGTA
2-G7_A                          CATTGACAGaagaAAAAATAAAAGCATTAGTa
2-G3_A                          CTTTGAcagaagaAAAAATAAAAGCATTAATA
2-G7_B                          CATTGACAGAAGAAAAAATAAAAGCATTAGTA
2-G3_B                          CTTTGACAGAAGAAAAAATAAAAGCATTAATA
2-G7_E                          CATTGACAGAAGAAAAAATAAAAGCATTAGTA
2-G3_E                          CTTTGACAGAAGAAAAAATAAAAGCATTAATA
2-G4_E                          CTTTGACAGAAGAAAAAATAAAAGCATTAATA
2-G7_F                          cattgacagaaGaaaaaATAAAagcattagta
2-G3_F                          ctttgacagaaGaaAAAATaaAagcattaaTA
2-G4_F                          ctttgacagAAGAAAAAATAAAaGcattaaTA
```

2-G3_B

```
     780       785       790       795       800       805
     113       118       123       128       133       138
con  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
edt  A A A T A A A A G C A T T A A T A G A A A T T T G T A C A G
phd  A A A T A A A A G C A T T A A T A G A A A T T T G T A C A G
```

2-G7_E

```
     780       785       790       795       800       805
     117       112       107       102       97        92
con  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
edt  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
phd  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
```

2-G7_B

```
     780       785       790       795       800       805
     126       131       136       141       146       151
con  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
edt  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
phd  A A A T A A A A G C A T T A G T A G A A A T T T G T A C A G
```